AD-A050 828 CENTER FOR APPLIED LINGUISTICS ARLINGTON VA F/G 5/7
SOCIOLINGUISTIC AND MEASUREMENT CONSIDERATIONS FOR CONSTRUCTION--ETC(U)
DEC 77 R F BOLDT, M K LEVIN, M GRIFFIN DAHC15-73-C-0364
UNCLASSIFIED AFHRL-TR-77-65 NL

| OF |

AD
A050 828

END
DATE
FILMED
4-78
DDC

AFHRL-TR-77-65

AIR FORCE HUMAN RESOURCES LABORATORY

# SOCIOLINGUISTIC AND MEASUREMENT CONSIDERATIONS FOR CONSTRUCTION OF ARMED SERVICES SELECTION BATTERIES

By

R.F. Boldt
M.K. Levin
D.E. Powers
Educational Testing Service
Princeton, New Jersey 08540

M. Griffin
R.C. Troike
W. Wolfram
Center for Applied Linguistics
1611 North Kent Street
Arlington, Virginia 22209

Forrest R. Ratliff, LtCol, USAF

PERSONNEL RESEARCH DIVISION
Brooks Air Force Base, Texas 78235

December 1977
Final Report for Period October 1975 – June 1977

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

## NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by the Center for Applied Linguistics, 1611 North Kent Street, Arlington, Virginia 22209, under contract DAHC 15-73-C-0364, and Educational Testing Service, Princeton, New Jersey 08540, under contract F41609-75-C-0034, project 7719, with the Office of the Assistant Secretary of Defense (Manpower & Reserve Affairs), Directorate for Manpower Systems Evaluation, Washington, D.C. 20301, in association with HQ Air Force Human Resources Laboratory (AFSC) (Personnel Research Division), Brooks Air Force Base, Texas 78235. Lt Col Forrest R. Ratliff, Chief, Manpower Development and Evaluation Branch (AFHRL), was the contract monitor.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved for publication.

LELAND D. BROKAW, Technical Director
Personnel Research Division

DAN D. FULGHAM, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
AFHRL-TR-77-65
OASD (M&RA) MR 74-12

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

9

4. TITLE (and Subtitle)

6

[redacted]

SOCIOLINGUISTIC AND MEASUREMENT CONSIDERATIONS FOR CONSTRUCTION OF ARMED SERVICES SELECTION BATTERIES.

5. TYPE OF REPORT & PERIOD COVERED
Final rept.
Oct[redacted]75 – Jun[redacted]77

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

10

R.F. Boldt
M.K. Levin
[redacted]
M. Griffin
R.C. Troike
W. Wolfram
[redacted]

8. CONTRACT OR GRANT NUMBER(s)

15

DAHC 15-73-C-0364
F41609-75-C-0034

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Center for Applied Linguistics ✓
1611 North Kent Street
Arlington, Virginia 22209

Educational Testing Service
Princeton, New Jersey 08540

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

16

62703F
7719 14

17

11. CONTROLLING OFFICE NAME AND ADDRESS

Office of the Assistant Secretary of Defense
(Manpower & Reserve Affairs), Directorate for
Manpower Systems Evaluation, Washington, DC 20301

12. REPORT DATE

11

Dec[redacted]77

13. NUMBER OF PAGES

42 p.

12

14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*

HQ Air Force Human Resources Laboratory (AFSC) ✓
Brooks Air Force Base, Texas 78235

15. SECURITY CLASS. *(of this report)*

Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

18

19

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

AFHRL, OASD/MRA

TR-77-65, MR-74-12

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

aptitude test
linguistics
test bias
test items

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The objective of this study is to identify potential sources of linguistics bias in Armed Forces aptitude tests. General aspects of a sociolinguistic perspective are set forth as a basis for investigating the use of language in tests. Linguistic interference is investigated for three different aspects of language usage in tests: directions; word problems, as seen in tests for arithmetic reasoning and automotive information; and word knowledge. For each of the main areas of investigation procedures for verification and experimentation are suggested, and further questions are explored. The conclusion summarizes specific considerations that should be given to sociolinguistic aspects of aptitude tests and suggests ways in which this analysis may be followed up by test designers and test interpreters.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

## PREFACE

This report is, in part, based on work begun by personnel of the Center for Applied Linguistics, under contract DAHC 15-73-C-0364. It was a theoretical study which had as its aim the identification of potential sources of linguistic bias in armed services selection aptitude batteries, particularly as these might affect the performance of members of ethnic minority and lower socioeconomic class groups. The study was based on the extensive body of linguistic writing and research data available on social and regional varieties of English spoken in the United States, and on the considerable amount of research in sociolinguistics and semantics underway at the time the contract was executed. Upon completion of that report, the Educational Testing Service, under contract F41609-75-C-0034, undertook to combine measurement considerations with materials produced by the Center in the earlier effort, resulting in the present report.

## Table of Contents

Table of Contents (Continued)

## 1. Introduction

Historically speaking, aptitude testing has been a major factor in manpower management since World War I, when the first large scale use of aptitude testing helped to mobilize military personnel. Since that time, the measurement of aptitudes has occupied a central position in such activities as personal counseling, educational planning, vocational training, and career and academic selection and placement. Tests have received extremely wide use in the selection and placement of applicants by employers, college admissions officers, recruiters, administrators, and job supervisors. When used for these purposes, tests are intended to benefit both institutions and individuals. The benefit to the institution accrues from the possibility of improved accuracy of selection, i.e., minimizing the number of applicants selected or placed who will subsequently fail to perform adequately. Thus, the institution is less likely to waste valuable resources to train individuals who are not likely to benefit from them. Similarly, individuals are thought to benefit, in that those whose probability of adequate performance is not great are not admitted, thus minimizing unproductive effort and resources by these individuals and also sparing them the personal trauma of failure.

However, while management may see the use of tests as an efficient way to channel talent, others often view the "gatekeeping" function of tests as a barrier to economic and social advancement. In the latter view, tests are threatening to those required to take them and a deterrent to the upward mobility of those whose performance on them is noncompetitive. In a high unemployment economy, job availability is likely to be restricted to those having even higher test performance. Thus, the visibility of tests, and perhaps the hostility toward them, is more prevalent (e.g., Byham & Spitzer, 1971, pp. 14-38, Griggs vs. Duke Power Co., 1971).

Test developers have the responsibility of ensuring their measurement instruments function as barriers to those unlikely to succeed in the selected tasks rather than those merely socioeconomically different from a normative group. Identifying potential sources of test bias and prescribing remedies is still an open issue among test developers. This report reviews the basic sources of test discrimination against minority ethnic or cultural subgroups, identifies sociolinguistic bias as an issue receiving little attention, proceeds to develop and explore a method for identifying sociolinguistic bias in tests, and then provides general guidelines for construction of selection batteries for use by the armed services.

## 2. Bias in Measurement

Dissatisfaction with tests is particularly great when it is noted that certain groups are consistently less successful: Some ethnic groups do better than others on tests of verbal ability (Anastasi, 1958, pp. 505-571); women are said to be handicapped on tests that require experiences more commonly available to males (Tyler, 1965, pp. 243-251). Blacks, women, and those for whom English is a second language all compete increasingly and visibly for jobs and professional standards set by the traditional job-holders of America--White men in appropriate age ranges (U.S. Department of Labor, 1973). Given this situation, it is reasonable to ask whether a low score truly forecasts low performance and whether the score difference is relevant to the purpose for which the person is to be employed. Furthermore, it is important to ascertain whether it is some temporary and easily remedied disadvantage of minority groups that accounts for the low scores that effectively exclude them from sought-after positions.

Large between-group differences in aptitude test performance have been noted for more than 70 years (Cronbach, 1975), and the source of these differences has been a topic of debate for nearly as long. However, only within the last decade has the relationship of group membership to aptitude measurement become a legal and social issue. Recently, the controversy has captured the attention of an increasing number of measurement experts who are directing careful thought and considerable effort to the problem.

## 2.1. Factors Contributing to the Definitions of Test Bias

An important assumption often made in interpreting test scores is that given reasonably comparable exposure to the culture, differences in performing reflect past differences in response to that culture. Furthermore, it seems reasonable to expect these differences to continue and to influence future job performance (Canady, 1971, pp. 89-101; Samuda, 1975, pp. 42-50). The premise of comparable exposure to a culture, however, may be untenable. In fact, there are those (e.g., Samuda, 1975, pp. 63-100) who believe that different groups (men and women, for example) are actually exposed to different cultures. The appropriate question is whether the resulting group differences in test scores are relevant to job performance. These differences may or may not properly reflect subsequent job performance, depending on a wide range of circumstances. Further studies relating group differences in test scores to on-the-job performance (e.g., Bray, 1972; Campbell, Pike, & Flaugher, 1969) are clearly needed.

The objective identification of test bias parameters requires consideration from more than a purely psychometric perspective. An early effort undertaken by an American Psychological Association (APA) task force (1969) to identify and define sources of bias in employment practices attempted to consider all aspects of the employee selection and promotion processes. These aspects include reception facilities, employer attitudes, aptitude testing, interview protocols, biographical data, and performance evaluation methods. The basic concern was the possibility of inadvertently introducing bias at various stages of the process, from the preliminary screening by the receptionist to the final decision made by the personnel director.

The basic recommendation made was that validation of objective data should be undertaken whenever possible to ensure that the information needed to make personnel decisions is both available and appropriate. The conclusion reached was that statistical validity as it affects the evaluation instruments is the most important factor in determining the presence of bias in the selection process. Thus, selection for employment or promotion should be made on the basis of as many objective, valid indicators as possible.

A number of court cases (Ruch, 1972) have provided quasi-legal descriptions of factors that may define test bias. Cases have included (1) those in which the prediction equation observed for minority groups is different from the equation computed for the general sample on which the test was validated and (2) those in which the percentage disqualified by the test is larger for minority groups than for the general validation sample. In one view, the existence of differences between the mean test scores of racial or ethnic groups (leading to different proportions being selected) is prima facie evidence of bias. In this view, the burden of proof is on the user to establish the validity of the predictor. A more recent Supreme Court decision (Washington vs. Davis, 1976) denies that prima facie evidence can be established merely on the basis of differentials in hiring rates (which may be associated with differences in test performance).

Cleary et al. (1975) have examined the assumptions and technical problems related to the use of aptitude measures in personnel decisions, making special reference to those aspects of test bias and fairness addressing test misuse, test score misinterpretation, and the measurement of multiple skills. They view the issue of fairness--which generally pertains to test use, not test content--as a problem common to both minority groups and the general population. The concept of fairness depends upon a number of factors, the major one being the responsible professional's knowledge of the strengths and weaknesses of the test and the appropriateness of particular applications. In this view, both bias and fairness are more strongly related to predictive (criterion-related) validity than to any other factor: The higher the validity, the more fair the test (or other measure). This statement also holds true when separate regression equations are generated to accommodate two or more groups in the population tested. Cleary et al. (1975) and Reilly (1973) describe situations in which over- or under-prediction results from an artifact of the population distribution:

when two groups can be assumed to come from the same general bivariate population, the predicted performance using a common regression line can be expected to result in over-prediction for the group at the bottom of the distribution when compared with prediction resulting from a separate equation computed for that group. Conversely, the performance of those at the top of the distribution will be under-predicted to some extent. Thus, if some identifiable group occupies a particular area at either end of the distribution of a sample sharing a common prediction equation, there will be a tendency to under- or over-predict performance, depending upon its rank in the distribution. Flaugher (1974) substantiates this fact, citing a number of studies in which the predicted performance of minority group members was better than their actual performance when a regression equation based on all groups was used.

Other definitions of test bias have been advanced by Thorndike (1971) and Cole (1973), among others. Thorndike indicates that even when validities are equal, tests may be unfair to lower scoring groups in the sense that the proportion who qualified on the test can be smaller than the proportion qualified on the job.

The use of the proportion who qualified versus the proportion who would succeed on the job seems to be a reasonable standard for determining the presence of bias. However, Cole (1973) advances the view that given one member of the majority group and one member of a minority group, both of whom would succeed if selected, fairness requires that each have the same probability of being selected.

It should be noted that these models of bias, including the purely statistical models, contradict each other in particular cases. In fact, Petersen and Novick (1976) point out that only two of the seven models they reviewed were internally consistent with respect to their logical converses. Cronbach (1976) suggests that, at the least, psychometrics can help lawyers and philosophers to "put more substantial arguments behind competing rules for obtaining equity" (p. 41).

## 2.2. Proposed Remedies for Bias

Three remedies for bias that have been suggested are (1) the elimination of testing, (2) the differential interpretation of test scores for different groups, and (3) purging the tests of sources of bias. The first remedy has been suggested in equal opportunity guidelines (EEOC, 1970). These guidelines imply that testing is inappropriate when the following conditions prevail:

(1) Validity data are neither available nor being collected.

(2) Promotion or selection procedures have adversely affected minority groups.

Fortunately, the tests used by the armed services have, in general, been subject to good validity research. The availability of many incumbents has permitted repeated validation in a variety of circumstances. The only apparent insufficiency--one that is universally common to validity research in all sectors--is the reliance on success in training, instead of on-the-job performance, as the criterion of success. However, adequate on-the-job performance measures generally do not exist, and training success may be more important since inability to complete training removes the opportunity for on-the-job performance.

The second remedy, differential interpretation of test scores, might be achieved by adjusting the scores of minority group members who are adversely affected by test use. An equivalent procedure involves making qualifying scores contingent on group membership. Other related procedures have been suggested also (Cole, 1973; Einhorn & Bass, 1971; Guion, 1966; Petersen & Novick, 1976). In practice, these procedures have often been used by universities wanting diversity in their student bodies. The modification of admissions standards for minority group members has on several occasions, however, resulted in legal action against universities (e.g., Bakke vs. Regents of the University of California, 1975; Ginger, 1974). The ethical issues involved in implementing different personnel processing procedures for different population subgroups are complex (Anastasi, 1968, pp. 280-286; Darlington, 1971; Kirkpatrick, Ewen, Barret, & Katzell, 1968, pp. 3-12).

The third alternative approach that has been attempted is the development of so-called culture free or culture fair tests that are valid predictors of job performance. The logical consequence of this concept--culture fairness--is that the average score of each subgroup will be the same. However, no such content has yet been found that will yield this result. Furthermore, the record to date strongly suggests that the search for completely culture fair content is not a promising activity (Anastasi, 1968, pp. 280-286; Dyer, 1960; Lorge, 1953, pp. 76-83; Tannenbaum, 1965, pp. 721-723). While complete culture fairness may not be probable, limiting sources of bias such as language usage may limit cultural bias in tests which are otherwise valid instruments.

## 3. Rationale for Investigating the Application of Sociolinguistic Principles to Testing

Because of its size, the military establishment is dependent on easily administered assessment devices for the evaluation, selection, and placement of personnel, particularly enlisted personnel. The devices used, and indeed massively used, are group administered, multiple-choice, objective, machine scored aptitude tests. Indeed, the advantages of such tests are so apparent that their use has also become pervasive in American industry and education.

All youths who seek entry as enlisted personnèl into any military service take initial selection and subsequent classification test batteries. The influence of the batteries is obvious: The strengths and weaknesses of military personnel tests affect the careers of a large segment of American youth. The development of techniques that improve the objectivity of military testing by reducing inadvertent variance due to linguistic structure or other unintentional complexities should have potential application to aptitude testing in general.

The present paper suggests that the developing body of sociolinguistic research might lead to the formulation of principles that could be used systematically to improve the language aspects of tests. The tactic adopted in the present work was for professional sociolinguists to analyze a sample of existing cognitive test material, identifying possible problems and seeking to determine the feasibility of formalizing sociolinguistic principles of test item language. At a later time, use of the resulting principles might help avoid language problems in future development of armed services selection batteries. The principles developed in this paper, however, should not be uncritically accepted and applied without rigorous investigation to determine effects on test reliability and validity in the test-taking population in general and in ethnic subgroups in particular.

The sections immediately following present some ideas about the potential contributions of sociolinguists to test construction. The major purpose of this effort is to provide a theoretical analysis useful in assessing the feasibility of applying linguistic concepts to testing. Mentioned are several approaches to (1) the systematic formulation of principles heretofore only informally stated and applied and (2) the identification and adoption of new principles of test construction.

## 4. Sociolinguistics

How is sociolinguistic research relevant to the construction and interpretation of tests?

In the past 40 years, a considerable body of research has accumulated on the varieties of American English. Such language differences reflect differences in the composition of society. Clearly age, class, ethnic group, sex, and geographical location all condition the language of a particular individual. This conditioning is, in turn, affected by the setting and purpose of any given language exchange. The nature and variety of American English that an individual employs and the facility with which particular varieties are used are functions of the user's socialization and personal history.

It should be noted that each variety of American English has its own degree of appropriateness to a particular situation. Each of the several ways of inviting someone to dinner ("Have you eaten yet?", "Do you want to eat?", "Are you hungry?", "Can you stay for dinner?", "You are cordially invited to join us for dinner," and "D'j'eet?") is appropriate for a given occasion. In addition, there are levels and kinds of language appropriate for spoken, as well as for written, language. Such a view is contrary to earlier judgments in which language was presented in terms of a simple dichotomy, the correct versus the incorrect. The more recent view rejects a single hierarchy of language levels--the kind of ladder that places the formal or literary at the top, the informal and colloquial in the middle, and the vulgar or illiterate at the bottom. Rather, it recognizes such categories as familiar and formal language as appropriate functional varieties.

The pluralistic nature of our social and educational structure seems almost to defy language classification. Clearly, a "standard"/ nonstandard" dichotomy does not seem adequate to capture the richness of a multidimensional language like contemporary English, nor does the value judgment implicit in such a dichotomy seem warranted. Nonetheless, it is true that those varieties of American English most often used to communicate formally in public settings, or to converse with non-intimates, lie at one end of a continuum. At the other end are those "nonstandard" varieties, which are used in less formal communication among intimates. Type of usage is also correlated with the educational background of the speaker, with more educated speakers tending to prefer the formal, standard variety. Informal or nonstandard usage by educated speakers would be placed near the middle of the continuum.

The language used in most tests is drawn almost entirely from the formal range of the spectrum. Furthermore, test language tends to reflect written rather than spoken usage. In particular, this variety--formal written--involves the use of complex sentence structures and vocabulary elements rarely found in the spoken language. But test takers differ with respect to previous exposure to formal standard language. Those who in their social environment have had less exposure to this variety will tend to have correspondingly less facility in speaking, reading, and writing it. This situation does not imply that the cognitive capacities of such speakers are limited. Indeed, the virtuosity exhibited by some individuals in their use of nonstandard language forms requires a variety of linguistic skills.

A hypothesis advanced in this paper is that the less exposure an individual has had to the language typically used in tests, the greater will be the linguistic difficulty encountered in taking the test. One would therefore expect the level of linguistic difficulty to be greater for those who typically employ nonstandard varieties of English or who come from environments where English is not the primary language. To the extent that these individuals are able to use the language of their own environments effectively, one would expect effective communication in new situations when given the opportunity to learn the linguistic demands of these situations and to practice skills needed to meet these demands.

Sociolinguistics, then, deals with the particularities of the interaction of language type and social experience. The evaluation of language correctness and the prescription of linguistic etiquette, however, are not proper functions of sociolinguistics. As a social science, sociolinguistics does aspire to a systematic understanding of the interactions between subculture, language variety, and language comprehension. It is anticipated that the application of sociolinguistic analysis and research will provide another perspective on some of the problems associated with the language of testing.

The present report does not promise a comprehensive treatment of testing problems from the point of view of sociolinguistics. Its purpose is to show by examples how a sociolinguistic application might be approached. An obsolete military selection test battery will be used as a representative and illustrative example. Accordingly, the discussion focusses on several areas in which language-related concerns are appropriate to test construction, administration, and interpretation. The ensuing discussion includes:

1. An examination of potential nonskill-related difficulties arising from language differences.

2. A consideration of test directions from a socio-linguistic viewpoint.

3. A statement of four sociolinguistic principles for evaluating test items and directions.

4. A critique of the synonyms item type.

The use of this strategy is not intended to convey a negative image of military tests. In fact, the relatively minor violations of principles in the test items chosen to illustrate points makes our examples seem at times somewhat labored. Many of the principles, therefore, might be more properly applied to tests and items containing more flagrant violations.

## 5. The Language of Directions

In any test battery, it is important that the test directions establish a common frame of reference for all the test takers. Only then can differences in individual performance be attributed to differences in the skill tested rather than to inadequate test directions. Orally administered directions are the information-bearing test elements for which it is easiest to infer equal examinee exposure. But, in spite of oral directions and the numerous pieces of clarifying information they convey, the assumption that the directions establish a common baseline should be seriously examined.

Since directions also serve as introduction to the test, some attention must also be focused on the setting and the atmosphere they create. Both of these conditions should convey the intention to be reasonable and helpful.

### 5.1. Read and Listen

The directions of the example tests were presented in two language modalities: the visual (written directions) and the aural (directions read aloud). Almost all directions are read aloud by the test supervisor to compensate for possible deficiencies in examinees' reading ability. This strategy is needed to ensure comprehension of the information by all participants because the general directions, as well as those in separate subtests, include fairly long and detailed passages. In fact, they were longer and more detailed than any of the test items.

The variety of English with which the examinee is familiar may well condition his ability to understand another variety. Examinees who have reading difficulties may also be relatively unused to reading <u>or</u> hearing formal English of the kind found in the sample tests. In this sense, the test gives an advantage to those social, economic, or ethnic subgroups who are comfortable with the type of language used in the test. Although it is not feasible to develop directions to which every examinee is accustomed, there are a number of language modifications that might be helpful. Some of these are given below; others are discussed under the principles presented in Section 7.

First of all, the examiner might be given more leeway in helping those who do not understand what they have heard. Indeed, the initial instructions in the example test strongly suggest that this should be done. The examinee reads: "Listen carefully to all directions. If they are

[1] Since there is a relatively common problem of being too explicit in communication events, achieving clarity is not as simple a matter as may be assumed. Giving more information than is necessary or giving it more often than is necessary violates Grice's (1967) Principle of Cooperation (i.e., that the language used follows the accepted purpose or direction of the language exchange in which one is engaged).

not perfectly clear, raise your hand. It is very important that you understand all the directions thoroughly." This instruction leads the examinee to expect that a request for clarification will be met with an additional explanation. However, if a question is raised, the administrator has been instructed to answer it only by reading the instructions, a procedure which may not be adequate if the problem is one of comprehension rather than hearing. Perhaps a set of alternative responses to frequently asked questions could be developed and furnished to test administrators.

## 5.2. Patterns of Repetition

Four information presentation patterns are found in the test battery. Some information is repeated on almost every page, some is reiterated for each subtest, and some is found only at the beginning of the battery. Other information is specific to some, but not all, subtests. The reasons for these different patterns of repetition are not immediately obvious. Regardless of their purpose, however, their value to non-standard speakers deserves examination, especially since they are stated in formal, standard styles.

Inconsistent patterns of repetition can seriously mislead the examinee. For example, at several points in the test the examinee is urged to work quickly but accurately. In the first subtest, this instruction is expanded with information about a 7-minute time limit. However, nowhere else in the battery is time mentioned. The examinee might, therefore, be led to assume that since no time limit is mentioned for the second subtest, none will be applied. This assumption is clearly inappropriate in light of the 10-minute time limit that is imposed on this test. The principle illustrated here is that when information is given, it sets up an *expectation or response set*. In order to avoid unwarranted conclusions by the examinee, directions should be such that all repetition is symmetric. Any changes in test requirements should by preceded by explicit instructions appropriate to these new requirements.

## 5.3. The Supervisor's Delivery

The use of emphasis and negative imperatives to ensure clarity is valuable but potentially risky. Obviously, the directions should be as helpful as possible in setting the tone of the examination situation. Emphasizing negatives and placing stress on particular words in a sentence, however, *may result in an irritating, unnecessarily authoritarian delivery*. Negative imperatives were frequently used in the test battery to repeat information first presented as a direct imperative. As such, they were probably a necessary expansion. In general, the stressed elements in directions to examinees conform to patterns of stress assignments found in the language as a whole (Bolinger, 1962, Crystal & Quirk, 1964; Pike, 1945). However, the assignment of stress in the directions read by test administrators is sometimes inappropriate in terms of normal language usage and may have undesirable effects. Stressing the last part of a compound in a sentence with normal falling intonation is unusual and distracting; yet it is required in the initial instruction given to each of the subtests (e.g., "Turn the page and BEGIN!"). The test administrator is also required to stress a one-word sentence ("STOP!") at the end of each subtest. Such distortions of normal stress patterns invite the administrator to shout in order to achieve the desired effect. In addition, the stressing of "any" in the last test ("Do not go back and work on ANY question in ANY of the other tests.") may be interpreted as a threat, instead of a simple order, by some of the more anxious examinees (Green, 1973; Sadock, 1972).

Directions could be easily rewritten to mitigate the potentially authoritarian tone produced by these stress patterns. Telling examinees to "BEGIN WORK" or to "STOP WORK" produces a more natural, less threatening intonation. In summary, the principle invoked here is that any distortion of normal speech in the test situation may be disconcerting to the test taker and should be avoided wherever possible. The use of a specific variety of English may in and of itself present difficulties for the test taker and, further, distortions of normal language patterns may create what appears to be a hostile environment. Insofar as these factors interfere with an accurate assessment of what is being tested <u>or</u> produce unnecessary antagonism toward the agency sponsoring the testing, they should be modified.

## 6. <u>Cultural Considerations</u>

The most subtle potential for test bias rests in the unstated assumptions, both social and linguistic, of the test constructor. Since these assumptions concern language or cultural matters regarded as inherently natural, self-explanatory, and completely obvious, the measurement expert may be hard pressed to recognize them as matters requiring attention. The linguistic example given below highlights the problem by illustrating a language feature that the native speaker would probably never question. Instead, he might assume that all languages are functionally equivalent, that they operate within the same frame of reference and make the same kinds of distinctions.

An example of the kind of problem that poses difficulties for nonnative speakers (even those who have attained relative fluency in English) is the use of the article <u>a</u>. This article has both a generic reading (e.g., <u>A human brain is heavier at birth than is a frog brain. She is a Marilyn Monroe</u>.) and an indefinite, specific reading (e.g., <u>A man came into the store this morning</u>.) (Lawler, 1972). In many test items, an object or person is first introduced in the generic sense and later, when further information is added or requested, treated in a specific sense. This procedure is prevalent in tests and may be considered a characteristic trait of test language. For example, "a man came into the hardware store and bought a quart of paint. He also bought . . . The prices were . . . How much did he spend?" In some languages, this ambiguity of the article a does not exist; an examinee whose native language makes the distinction explicit might not automatically equate "a man" and "he," and so may be confused by the ambiguity in test items in English. The problems, which do not exist for those who speak only English--but may exist for others--can be ameliorated by substituting proper names or other specific designations for "a man."

More pervasive in the test battery, but more amenable to correction, are the cultural assumptions that condition what is the "best" answer to a given test question. These are most apparent in those subtests where objective criteria for determining correct answers are either unclear or unavailable. The following item, taken from a Word Knowledge Test, illustrates the point:

Potent means

A heavy
B royal
C powerful
D drunk

The examinee, asked to choose between "heavy" and "powerful" in finding a synonym for "potent," but who does not know that in formal English "heavy" could not mean "potent," is at a disadvantage, particularly if the word has that meaning in the examinee's own speech.

While the relatively minor defects in the particular items presented above may not be especially harmful, the point to be made is this: There are subtle differences in the structure of languages, both formal and informal, that create a potential for the inadvertent introduction of ambiguity--and possibly bias--to tests. Careful review of test content by thoughtful test constructors and/or language experts could probably eliminate most major problems.

### 6.1. Values Specific to the Majority Culture

The fact that society places a high value on verbal ability is not itself a problem; deciding which aspects of verbal ability are important, however, is a problem. The example tests' heavy dependence on vocabulary items reflecting an extremely formal style (Shall I inform him? Cross the road with caution.) implies that knowing words of this kind is of prime concern. In addition, the stimulus item is typically a more difficult word than the correct response. Proceeding through the Word Knowledge subtest, the examinee becomes increasingly aware of the examiner's tendency to use formal words as item stems and more common ones as alternative responses. Although this lack of symmetry may be perplexing to some examinees, it is actually intentional. The use of alternative responses that are more likely to be known by all examinees helps to ensure that incorrect responses result from unfamiliarity with stimulus words, and not with response alternatives.

In several instances, test items may penalize particular subgroups of the test-taking population. The word feat, meaning an accomplishment showing unusual skill, is illustrative of a particular type of defect. A Spanish speaking examinee misreading this word as fete (festival) or trying to relate it to a Spanish cognate may mistakenly choose the word celebration as the correct answer. This examinee appears, therefore, to be penalized by attempting to exercise a productive and useful bilingual skill. It is likely that this item may indeed fulfill the purpose for which it is intended--discriminating between examinees who know the word's meaning and those who do not. The point, however, is that, in the face of uncertainty, some feature of the examinee's language or culture may determine the attractiveness of alternate choices. The example given here suggests that a non-Spanish speaking examinee who does not know the meaning of the word fete might make a random choice, thereby having a 25% chance of correctly answering the item. Spanish speaking examinees, on the other hand, might more frequently employ the bilingual skill mentioned above, choosing a particular incorrect alternative, celebration, more often. When attempting to devise plausible alternatives or multiple-choice items, test item writers should exercise care in order to reduce the possibility that specific alternatives are not differentially attractive to those subgroups defined by common cultural or linguistic characteristics. Standard item analysis procedures could be used to empirically assess possible differences.

## 6.2. Other Particular Problems

Another potentially troublesome situation becomes apparent when one realizes that most words have several possible, sometimes divergent, meanings. The implications of multiple meanings can be shown by referring to four words found in the Word Knowledge subtest.

According to Webster's Third International Dictionary, the word ample is defined as:

1. Marked by extensive or more than adequate size, volume, space, or room.

2. Buxom, portly.

In light of these definitions, two of the alternative choices, fat and well-shaped, might be considered as defensible choices. Well-shaped might be chosen by an examinee whose subculture considers portliness to be a physically attractive quality.

Likewise, an archaic definition of scour ("beat, punish") recorded in the same dictionary might make the choice of whip acceptable. Similarly, one definition of sullen ("of a dull color, of somber hue") could possibly make two of the choices, grayish yellow and very dirty seem reasonable. A closely related problem is illustrated by an item testing the meaning of terse, defined in Webster's Third International Dictionary as "smoothly elegant: polished, refined" and "devoid of superfluity: brief, concise." Although the keyed response, pointed, is the best choice available for terse, it is not an obvious synonym for either of Webster's definitions.

Granted, the problems illustrated are not severe in the sample test, especially since the instructions direct the examinee to select the best answer. However, one must ask the question, "Do vocabulary items with these types of distractors represent the most effective approach to measuring vocabulary or verbal ability?" Are these kinds of word discriminations, which may in fact have a spurious attractiveness for some subgroups, the best choices which could be made if viewed from a semantic or linguistic perspective?

## 6.3. Errors of Omission

In constructing a test such as Arithmetic Reasoning, test writers typically use examples which they assume will reflect the everyday experiences of most examinees. In doing this, however, the tester may exclude useful material. It seems appropriate, therefore, to examine test materials to determine what the examiner may have omitted as he tried to select only common material.

The sample test's failure to reflect the diversity of the population taking the test illustrates the tendency for omission. Persons named in the test are called Tom, Bill, John, or Joe--typical white, middle-class names. The Puerto Rican or Mexican-American finds nothing in the test that acknowledges the existence of his culture. Women are conspicuously absent also, even in traditionally female situations such as purchasing food and clothing. This practice certainly avoids stereotyping but at the cost of ignoring women completely. Attention to such details might well lead to the inclusion of a greater variety of material--material that would produce a more appropriate balance of content with no sacrifice in clarity or reasonableness. Even minor revisions might have a beneficial psychological effect on minorities or cultural subgroups.

## 7. Formulation of Some Sociolinguistic Principles

As indicated in the Introduction, this report predicates the potential value of sociolinguistic principles formulated with test construction in mind. Because such principles are not readily apparent from the examination of the literature of either sociolinguistics or testing, active steps are required to bring the formulation about. To do this, specialists in various aspects of sociolinguistic study were directed to use their knowledge of different varieties of English and ethnic and minority value systems in order to predict potential sources of difficulty in the test battery. These specialists were chosen for their work on language differences in American English, including standard and nonstandard regional variations, and for their research experiences with the problems of non-native speakers. Their task was to explore the application of sociolinguistic principles to two of the sample subtests (Arithmetic Reasoning and Automotive Information) that rely on language to formulate individual test questions.

A judgmental analysis of these subtests indicated that four specific sociolinguistic principles are important both in describing areas in which minority examinees encounter difficulty and in suggesting remedial action to neutralize these difficulties.

### 7.1. The Principle of Pragmatics

The principle of pragmatics states that the values implied or stated in test items should be consistent with the values of the examinee. Mass testing procedures often assume that the item writer and the examinee understand an item within the same frame of reference. The test constructor cannot know the value systems of all the subpopulations who will take the test, but a sociolinguistic reviewer may be able to alert him to potential problem areas. An examiner sensitized in this manner could, presumably, avoid difficulties arising from differences between examinee values and those implied in test items--differences that usually arise from differences in the backgrounds of examiners and examinees. The examples below may help to clarify the differences in values that are likely to be enountered.

> An insurance policy costs $7.70 a month, or $85.00 a year. How much money will a person save each year by paying for a year's insurance at one time?
>
> A $ 5.00
>
> B $ 7.40
>
> C $ 8.40
>
> D $92.40

A man paid $150 for a set of 4 new tires. After using them for 10,000 miles and paying out $8 for repairs, he received $2 apiece for them toward a new set. How much per mile did he pay for the set of tires?

A $.0134
B $.0150
C $.0168
D $.0672

These items, dealing with buying insurance and calculating the cost per mile for tires driven over a long distance, presuppose familiarity with the allocation of financial resources. This assumption, however, is not necessarily realistic for examinees from low-income backgrounds. For example, low-income examinees typically experience situations in which credit buying is customary. Insufficient income often prevents the choice of any other type of payment, making decisions related to credit versus noncredit buying somewhat academic. The concept of long-range benefits, as invoked in the insurance item, may be completely foreign to the low-income examinee's economic frame of reference. To those who have internalized the value system of the impoverished, these items call for decisions that might be strange. A difficulty with strict application of the principle of pragmatics is that some values and experiences, as in work values, may be highly relevant to the demands of the job environment and hence be important to the validity of the test item. Care must be taken to evaluate critically both sides of the issue on an item by item basis. In summary, the principle of pragmatics suggests that test items should avoid posing situations that are uncharacteristic or atypical of the life styles of test takers, especially when these situations are experienced differentially by various examinee subgroups and are not criterion related.

### 7.2. <u>The Principle of Processing</u>

The principle of processing, reflecting the assumption that items can be categorized in terms of the language and reasoning processes they require, suggests that particular item categories, or subtests, should contain only items that require the same process(es). The term "processing" is related to the test taker's ability to respond appropriately to different types of information ordering. This entails dealing with situations in which the nature of the information given varies in several significant ways.

Several items in the Arithmetic Reasoning Test appear to require different combinations of information processing skills. Consider the following items:

Tom bought 108 pounds of nails.
If he gave 32 pounds to his
brother, how many pounds did he
have left?

A 140
B 86
C 76
D 72

This information presentation requires only a simple subtraction. The item can be answered without recourse to the answer choices.

On the other hand, an examinee must first consider the alternative choices in order to arrive at the expected answer for the following item:

An article that sells for $5.00 costs
a customer $5.10 when the sales tax
is included. What is the sales tax?

A 5%
B 3%
C 2%
D 1%

The correct response, if answered using only the information presented in the item stem, would be "10 cents," rather than 2% as required by the options. Here the examinee must rely on information given in the stimulus material and on the answer choices, since the question makes no mention of percentages. In addition to the simple calculation required, the test taker must also realize that an additional step, conversion to a percentage figure, is implicitly demanded. The discrepancy can be avoided by following the test construction practice of having a completely self-contained stem. In the above example, stating the question as, "what is the percentage of the sales tax" can solve the problem. Now the examinee can rely on the stem or stimulus material to arrive at the answer.

Still another set of information processing skills is needed to answer another type of Arithmetic Reasoning item.

Joe buys 9 shirts and pays $1 for a tie.
The total cost is the same as Bill spends
when he buys 4 shirts and pays $11 for a
hat. If all shirts cost the same, what
was the price of each shirt?

A $2
B $3
C $4
D $5

To answer this item type correctly, the examinee must set up and solve algebraic equations.

It is important to note that although all the various types of items require active calculation on the part of the test taker, they differ in the kind of information requested and the type of process required. In summary, it seems that a particular response set may be established by a series of items requiring similar information processing or reasoning skills. It is suggested that subsequent items should not require widely different skills, unless the test is designed to reflect the ability to select appropriate processing strategies. Although a test like Arithmetic Reasoning may have this purpose, there are other tests that do not. Care should be taken to ensure that when items differing with respect to type of reasoning processes required are included within a given subtest, the varied items were included by design and are necessary to the purpose of the test. For example, a varied sample of reasoning processes would be required in the case of summation scores where higher scores are intended to mean more ability/mastery of mathematical principles.

7.2.1. Too much information. In some items, the examinee will encounter a mismatch between the amount of information available and the amount needed to solve the problem. A test taker may anticipate that all the information given in a problem is to be used in its solution, only to find out later that some of it is irrelevant. This situation may or may not be desirable depending on the tester's purpose. If the purpose is to assess the examinee's ability to ignore irrelevant information, including such information is quite appropriate and, in fact, necessary. This practice is commonly used in the development of the so-called data sufficiency items found in a number of well-known tests.

If, however, the tester's purpose is to assess the ability of the examinee to reason from relevant information, then it seems desirable to include only information required to solve the problem. Consider the following item:

> Two cars started from the same town at the same time. One car traveled 50 miles an hour for 4 hours. The other car traveled 60 miles an hour for 8 hours. How many miles farther did the second car travel?
>
> A 10
> B 40
> C 200
> D 280

Giving information about the starting time of the two cars leads the examinee to expect that the solution will in some way involve the arrival and departure times of the cars. However, the information given in the first sentence of the item is unnecessary for the problem's solution; some would regard this information as completely extraneous. In essence, the inclusion of such irrelevant information violates a principle of language usage that Grice (1967) has labeled the Maxim of Relation, a principle assuming that only relevant information is given. Violation of this principle not only fails to meet basic test construction principles but the increased verbiage has particularly devastating effects on poor readers and normally poor test performers prevalent among many different socioeconomic groups. A sociolinguistic application of this principle to testing would suggest that considerable effort should be taken to avoid inclusion of irrelevant information in test items.

7.2.2. Insufficient information. In the example below, which deals with lump-sum versus monthly payments, it is possible from the way the facts are stated to suppose that the lump-sum figure and the monthly figure are equivalent, unless the test taker stops to calculate their relationship.

> An insurance policy costs \$7.70 a month,
> or \$85.00 a year. How much money will a
> person save each year by paying for a
> year's insurance at one time?
>
> A \$ 5.00
> B \$ 7.40
> C \$ 8.40
> D \$92.40

Nothing that is overtly stated makes it clear that the annual rate is less than the monthly rate, and test takers from low-income backgrounds are unlikely to be aware that such is usually the case. A simple rewording of the item would add to the verbal content but make it more acceptable.

The most serious problems of insufficient information involve those items that allow legitimate alternative tracks of reasoning and lead to answers which are scored as incorrect. For example:

> Gasoline costs 20 cents a gallon before
> taxes. There is a 20% road tax on each
> gallon of gas, as well as a 5% city tax
> and a 5% state tax. What is the total
> cost of 8 gallons of gasoline?
>
> A \$2.08
> B \$2.40
> C \$2.80
> D \$4.80

This item allows for the computation of taxes based either on an accelerated figure or on a constant base price. Using the accelerated approach, the examinee would take 20% of the base price (20 cents) and add the computed tax (4 cents) to the base price. Additional taxes would be applied to the new total at each step. Although using this accelerated procedure may not be strictly correct, the current use of the ever popular surcharge might make such a choice seem reasonable to many examinees. Since the item is intended to assess arithmetic reasoning, not specific knowledge of tax computations, the apparent ambiguity should probably be rectified by including additional information.

### 7.3. The Principle of Formality

This principle states that the greater the distance between the variety of English familiar to an individual and that used in a test, the greater will be the potential linguistic difficulty for the examinee. The problem is more serious when there are marked differences between the variety of language an individual speaks and the variety which he must read than when an individual's spoken usage more nearly approximates the written form. Nonstandard spoken language varieties are most characteristically employed by infrequent readers (who are often of lower socioeconomic class background) and in informal settings. Given that most tests are written in a relatively formal variety of standard English, the principle states that the level of linguistic difficulty would tend to be systematically greater for individuals from lower socioeconomic backgrounds and backgrounds where English is not the primary language than for those from middle-class backgrounds.

The type of language used in testing often has certain peculiarities that distinguish it from the language of everyday conversation and even from the formal standard English found in other types of writing. For the most part, these differences are in sentence structure and vocabulary choice, and they constitute probably the more serious and more correctible sources of potential bias in the example test battery. For example, a sentence like the following, not uncommon in standardized tests, would be relatively rare in spoken English:

> When measuring an unknown voltage with
> a voltmeter, the proper precaution to
> take is to start with the ...

No reduction in clarity or diminution of context would result from reworking this item to read as follows:

> In measuring a voltage with a voltmeter,
> you should be careful to start by ...

In this rewording, the vocabulary and the syntactical arrangement conform more closely to natural conversation, thereby eliminating the barrier resulting from unnecessary formality. The content of the question remains unaffected.

Throughout the test, more formal lexical items are consistently chosen over more familiar ones. Words like locate (instead of find), obtain (instead of get), fails to (for doesn't), and approximately (instead of about) all reflect such choices. Though certain other items may appear to be innocuous, further investigation reveals that there may be subtle shades of meaning involved which could lead to further misunderstanding by some test takers (Gordon & Lakoff, 1971; Green, 1973).

## 7.4. The Principle of Redundancy

The principle of redundancy states that the redundancy-reducing rules characteristic of written English may cause difficulty for examinees whose familiarity with formal written English is limited. These rules serve to reduce redundancy by deleting information that is identical to information previously stated, by converting relative clauses to more abbreviated constructions, and by introducing various references to previously mentioned material.

For example, the deletion of the preposition by in a sentence such as "Bill makes ten dollars a week (by) washing cars" makes the sentence slightly less clear (though perhaps more conversational). Similarly, the use of a reduced clause construction in reference to a container that weighs "1,200 pounds empty" is less clear than the full construction "1,200 pounds when it is empty." In other items, the kinds of reduction allowed by English grammar in comparative sentences may have been used to the potential disadvantage of some test takers. When reduction is applied to comparative sentences, ambiguity may be introduced and comprehension reduced. For example, a sentence such as "John has helped more people than Bill" is ambiguous. It can mean "John has helped more people than (just) Bill," or "John has helped more people than Bill has helped." It would be better to give the fuller form, "John has helped more people than Bill has helped," if that is the intended meaning.

The item below begins with a complex sentence to which a syntactic deletion rule called "gapping" has been applied.

> The running time of a movie is 1 1/2 hours, of the newsreel 10 minutes, of the cartoon 8 minutes, and of the coming attractions 7 minutes. At what time would the entire show be over if it began at 6:50 p.m.?
>
> A 8:05 p.m.
> B 8:25 p.m.
> C 8:55 p.m.
> D Some other time

Gapping allows redundant material to be deleted in a series of similar constructions after it has been stated in the first member of the series (e.g., "John ordered fish, and Bill (ordered) liver." Gapped sentences may be quite difficult to follow; a very substantial reduction in difficulty might be achieved in this item by giving the full ungapped form. In other instances, however, gapping may be effectively applied to reduce passage length. Inclusion of redundant material often helps slow readers, especially lip readers, and individuals less familiar with formal English to understand the content of the test items.

7.4.1. Reading level difficulty. In the above paragraphs, it will be noted that the proposed revisions involving redundancy-reducing rules quite often require an increase in the length of the sentence. Since some of the more traditional scales for measuring reading difficulty (such as the one proposed by Flesch (1951)) view reading difficulty as a function of sentence length and the number of polysyllabic words, one might question the effect of redundancy-reducing revisions on reading difficulty. We suggest that perhaps Flesch's conclusions are more relevant to some situations than others. For example, some item writers may employ a style relying on complicated grammatical constructions and difficult vocabulary. To these writers, Flesch's approach clearly offers a guideline for remedying their stylistic defects, especially when the audience is homogeneous and relatively proficient in the language used. But the enlisted military selection tests place demands on item writers that are much more rigorous, perhaps requiring other measures than those suggested by Flesch's. Among these other measures might be the principles suggested in the previous section.

## 8. Experimental Application of Sociolinguistic Principles to Word Problems

The development of new principles or constructs such as those evolved from a sociolinguistic context raise numerous questions concerning their utility, methods of application, the reliability or validity with which their elements can be discriminated, and perhaps their influence on increasing the clarity of meaning in written statements (test items). The lack of empirical data on these questions led to the performance of a small pilot study to observe basic rating characteristics, response patterns, and influence of type of subject matter on a rater's judgements. The three persons who assisted in the development of and were thoroughly familiar with the four principles, i.e., pragmatics, processing, formality, and redundancy, were requested to rate the items in two sub-tests of the sample tests.

The judges were asked to indicate whether or not specific terms violated the principles and, if so, which principles were violated. The analysis indicated that on one subtest judges agreed with each other reasonably well. They agreed upon (1) the items which violated sociolinguistic principles, (2) the severity of the violation, and (3) the particular principle involved. There was a noted lack of agreement, however, between the judges on the other subtest with very few indications by two of the judges of a violation of sociolinguistic principles.

The degree of relationship found between judges on one of the sub-tests suggests that the four principles can, with further experimental refinement, be used to identify potential sociolinguistic problems in test items.

## 8.1. Future Applications

A thorough application of sociolinguistic principles to test development would require a more extensive effort than the attempt made in the present study. It would entail the following steps: (1) a set of materials would be examined by sociolinguists, who would then formulate a set of principles and adequate rating scales for dealing with the language of tests; (2) the resulting principles would be applied to a new set of materials to produce tests free from the previously described defects; (3) unrevised, but otherwise identical tests would also be assembled, and the two sets of tests would be administered to random halves of a group of examinees. Differences in the test score performance of examinees under each condition would be noted and subsequently validated against a relevant criterion. These procedures should be repeated using different materials, groups, and types of subsequent validating performances. Different sociolinguistic experts could also be employed to develop different principles to be examined. Clearly, the number of possibilities would preclude an all-inclusive investigation. This should not, however, discourage more modest efforts.

## 9. The Word Knowledge Subtest, Synonyms

The Word Knowledge subtest is the only test in the example battery specifically intended to assess a verbal skill. If any of the tasks to be performed in this subtest are not related to word knowledge, then the content validity of the test might be questioned. For a sociolinguist, an attempt to establish content validity would entail framing a concept for the term "word knowledge" and then determining if the items satisfy the concept. An even more appropriate method would involve writing test specifications as implied by the concept. Since we must deal here with an existing test, the latter approach is not possible.

Doing well on the Word Knowledge subtest requires at least three qualities: the ability to read, a notion of meaning and synonymy, and a knowledge of a sufficient number of words tested. Other more subtle skills one might wish to test include:

1. Knowledge of syntactic constraints (i.e., knowing into what sentence structures particular words fit).

2. Knowledge of stylistic constraints (i.e., knowing for what linguistic and social settings particular words are appropriate.

3. Knowledge of semantic constraints (i.e., knowing with what other ideas particular words can be used).

4. Morphological information (i.e., knowledge of word origins and derivations).

5. Knowledge of relationships to other words.

6. Knowledge of the presuppositions implicit in words, and their implications.

7. Knowledge of the pronunciation and spelling of words.

The Word Knowledge subtest does not seem to demand all seven of the knowledges listed above, although each might be helpful. This suggests that there is no full assessment of the examinee's word knowledge, nor was one intended.

But there are problems encountered in the use of the synonymic form beyond the limitations previously described. One type of mismatch is set up in the directions in subtask 2 of the test where the candidate is asked to decide which choice "most nearly means the same" as the stem word; in an example the wording shift, incorrectly and unfortunately, to "means the same." Clearly the former more accurately reflects the task than the latter, since very few words are exact synonyms, though they may be judged approximately so. Mismatches also occur between stem words and correct alternatives; three of the most frequent kinds of such mismatches are given below.

## 9.1. Lack of Semantic Equivalence

In the Word Knowledge subtest, knowing which of the alternatives carries the same semantic content is very helpful. Experience teaches that one-to-one equivalence of this kind rarely, if ever, exists. Even though a limited set of experiences may yield the judgment that a pair of words are synonymous, only one relatively minor experience is needed to disprove the judgment. (See Binnick, 1971, 1972 and Lakoff (1972) for just such instances of disproof of synoymy.) Even in such a close pair as sweat/perspiration, the words are not equivalent in all situations; horses sweat, while people perspire. A man lives by the sweat (not the perspiration) of his brow. The differences are also apparent in humour triads such as: I am firm. You are obstinate. He is a pig-headed fool.

### 9.2. Scalarity

Language users often behave as if an implicit ranking procedure operates for many word pairs. Words that refer to approximately the same objects can differ in relative strength. In the following examples, for instance, a weaker word is used in the simple sentences. The assertions in these sentences are made stronger if the phrases in parentheses are added:

She's intelligent (, in fact, she's brilliant).

The children are happy. (What's more, they're ecstatic).

I'd say this land is pretty (, even beautiful).

Note that reversing the order of intelligent and brilliant, happy and ecstatic, and pretty and beautiful (that is, switching to a stronger first word) produces a particular type of verbal joke.

### 9.3. Generality

A second type of difference between the stimulus and response words concerns the distinction between the general and the particular. Related words, especially those that are mutually substitutable in at least some situations, can be ranked in two very general kinds of hierarchical structures (cf. Bever & Rosenbaum, 1970). The following sentence frames can be used to determine if either hierarchy relates to a given pair of words:

1. A ________ is a part of a ________.
2. A ________ is a kind of ________.

For words other than nouns, minor modifications of the frames will yield the correct judgments. The first blank in each frame will, of course, be filled by the less general term of a pair. For example, quiet-calm and blemish-defect are such pairs, the first item in each being contained within the hierarchy of the more general second item.

## 10. Perspective and Prospects

The foregoing sections have presented a number of sociolinguistic considerations about the use of language in test construction, and have raised a number of issues needing critical examination. The present section will review some of these issues from a psychometric perspective and then suggest steps that might lead to an appropriate use of sociolinguistic techniques in testing.

## 10.1. Perspective

In testing, as in many other areas in the social sciences, practice of the art is difficult because everyone considers himself an "expert." Therefore, there exist many commonly held beliefs that are unsupported, or indeed even contradicted, by evidence. Frequently this evidence is known by only a small group of researchers, while the belief is popularly accepted and widely held. A few such beliefs are presented and then qualified below.

Belief One -- Test language is unnecessarily difficult. If simpler language were used to pose questions, examinees unaccustomed to academic English would perform better. This contention has been tested by Bornstein and Chamberlain (1970) who, noting the difficulty of language in tests of social studies achievement, rewrote test items using simpler language. They found highly similar performances for the easy and hard language versions, a finding that is supported by a similar study (Livingston, 1973).

Belief Two -- Psychological tests are not fair to groups who achieve low average scores. This belief ignores the need to relate scores to job performance. The military services' extensive programs of research and development confirm that low scoring personnel may realistically be expected to perform less well on the job than high scoring personnel.

Belief Three -- Psychological tests may be valid for most people but are not related to the performance of minority group members. The proponents of this belief have been so influential that it is mentioned in the guidelines developed by the Equal Employment Opportunity Commission (Guidelines in Employee Selection Procedures, 1970), and, indeed, there may be groups for which the belief is true. The extensive research conducted to date, however, shows tests to be equally valid for minority and majority groups. Boehm (1972) and Schmidt et al. (1973) have surveyed the literature of validity differences for Blacks and Whites and have found that, except in a few studies characterized by small samples and inadequate controls, substantially lower validities for minority groups have not been demonstrated.

Belief Four -- People who are unfamiliar with tests are at a disadvantage. A little coaching on test taking would improve their scores. If this belief were true and if score gains were reliable, many examinees would be expected to benefit from coaching. Unfortunately, such is not the case. In three studies sponsored by the College Entrance Examination Board (Angoff, 1971), coaching was attempted to increase test scores. These attempts, made at a high-prestige private institution (Dyer, 1953), at a public institution (French & Dear, 1959), and at a rural school in a depressed area (Roberts & Oppenheim, 1966), were not successful in raising total test scores. It is currently felt, however, that coaching might help reduce anxiety for some examinees, and might improve performance on certain specialized item types. Any such score gains, however, are expected to be neither large nor pervasive.

Although the existing evidence does not support these beliefs, some of them are undoubtedly implicitly involved in certain of the issues raised in the preceding sections. In evaluating the discussion in these sections, therefore, the following considerations should be kept in mind:

1. The sociolinguistic principles and evaluations developed in this report result from a first attempt on a limited amount of material and should not be judged as a finished or final example of scientific application.

2. The principles and evaluations are not to be regarded as universally true, but applicable only in certain situations.

3. The principles and evaluations are only a small part of the contribution that might eventually be made by the application of sociolinguistics to testing.

4. The principles and evaluations are not uniquely the property of sociolinguists; many of the items identified as defective by sociolinguists could also have been so identified by test constructors for similar reasons.

The systematic development and application of sociolinguistic principles to testing will require much more precise formulation and testing than has occurred to date. Some steps in this direction are suggested below.

## 10.2. Prospects

The application of sociolinguistic principles to test construction would occur in setting test specifications, writing and reviewing tests and items, and developing interpretive materials. The actual principles should, to the extent possible, be formalized, and the effectiveness of their application should be researched. In light of the plethora of beliefs that have been substantiated only occasionally, research is particularly important in applications dealing with population subgroups.

10.2.1. Specifications. Test or test battery construction requires adequate test specifications, regardless of the purpose and context of testing. In some situations, such as academic selection, there have been literally hundreds, perhaps thousands, of validity studies. The most effective predictors are well known and can be specified in advance. But many situations encountered in the military services require the early identification of those who will perform well on some relatively unstudied task. In this case, a variety of item types must be tried to define those appropriate for use in a selection battery.

Test and item specifications should include item type (e.g., analogies, antonyms), content (e.g., verbal ability, automotive information), statistical specifications (e.g., percent passing each item and minimum acceptable item-test correlations), and other important factors such as the number of items, testing time, physical format, and choice of directions. In choosing an existing set of directions or in writing new ones, a tester could usefully apply sociolinguistic principles to make the following decisions: what kind of directions (oral or written) to use, what level of language is appropriate, how much flexibility should be given to test administrators, how to use imperatives in giving instructions, and what level of previous exposure to testing to assume for various groups of examinees. At present, decisions with respect to these various aspects of directions are based primarily on logistical convenience, on existing standard practice, and on the assumption that identical procedures accomplish equal exposure. Better specifications or better support for the existing specifications for directions, as well as other aspects of tests, might result from a sound research program.

10.2.2. Item writing. The item writer could have available a set of research results and principles that could be used in formulating items. Some decisions regarding item format would, of course, have been made when test specifications were established. For example, the use of extraneous or insufficient information would be a matter of choice in some item types, such as those in which the examinee must determine which of several given reasons are sufficient to establish a stated conclusion. But inadvertent extraneous information might also be usefully included in arithmetic items. It would, therefore, be helpful to an item writer to know when he could legitimately complicate the problem posed by the item, and when he could be handicapping a group whose subcultural expectation of test taking is that all of the information given must be used. The item writer should have at hand some indication of the effectiveness of attempts to remove such expectations through modification of directions.

The item writer must also confront directly the problem of writing difficult items, items in which the difficulty arises from the nature of the problem posed, not from the language in which the problem is stated. Perhaps sociolinguistic research could lead to a separation of language difficulty and problem difficulty so that one could learn to pose hard problems in easy language.

10.2.3. Item review. As with many other creative acts, the writing of test items can proceed in two steps: in the first, the central idea of the item is conceived and put on paper; in the second the rough idea is developed and polished. The principle of pragmatics is one that could be applied in this second stage, since it implies that an otherwise appropriate problem could discriminate unfairly if put in the wrong context. The item reviewer should, therefore, be relatively free to consider background information related to the language and culture of ethnic, religious, and sex groups. He should also be attuned to the possible implications that such information has for test items. Eventually, a checklist of principles could be developed for use in evaluating each item for linguistic and cultural defects.

10.2.4. Test review. After assembling the test, the items and directions should be examined. At this step, the principle of processing would be applicable since it deals with items in combination. This principle emphasizes that answering items having similar content may require different logical processes. The principle, as stated by the sociolinguist, suggests that processes should not be mixed. While the tester might not be averse to mixing such processes, he would undoubtedly prefer that it be done intentionally. One aspect of the test review, then, would be to check and evaluate possible contradictions of the principle of processing.

10.2.5. Pretesting. Good testing practice requires that new items be administered on a trial basis so that unsuspected defects can be noted. Some major testing organizations conduct programs of pretesting and maintain test files that contain a record of each item's statistical performance. In light of the previous discussion, it seems desirable to keep the results of statistical analyses of items on population subgroups. It should be emphasized that group by item interactions, not overall group differences, would be the most informative indicator of the quality of items. Angoff and Ford (1973) have long asserted that such comparisons of item difficulty in groups could be used to identify particularly troublesome items. For example, certain tool knowledge items might be more difficult, on the average, for women than for men, since some of the tools mentioned are seldom found outside factories, which are traditionally men's domain. More common tools likely to be found in home workshops might be more equally recognized by men and women.

10.3. Research. It seems likely that the full benefits of sociolinguistics in testing will require an extended period of development, application, and evaluation of principles and information. Its organization, mission, and access to diverse populations makes the military service better suited to carry out such a program than most other establishments. Military personnel research in the application of sociolinguistics to testing could produce results that have value not only to the military establishment but to industrial and educational organizations. This, of course, assumes that the discipline has the potential and that research results are disseminated through appropriate professional journals.

Although a complete formulation of a research program of this nature is beyond the scope of the present paper, some aspects of such a program are given below.

10.3.1. Some research topics. Developing a research program that is both comprehensive and relevant to the requirements of the military establishment goes beyond the resources and scope of this paper, but some topics can be listed. Clearly, the research required to implement the development and application of sociolinguistic principles to the areas identified in the previous section must address a number of issues. Some of the areas that sociolinguists have felt might be usefully investigated are listed below:

- the inclusion of extraneous information in reasoning items,
- the degree to which the context of reasoning problems is appropriate to specific subcultures,
- the use of redundant language in test items and directions,
- the changes in the types of information processing that are required by certain items,
- the use of various algorithm-specific directions on coding speed performance,
- the modification of statements of purpose found in the directions,
- variations in the degree of flexibility given to test administrators, and
- variations in the level of difficulty of test language (e.g., extensions of the Bornstein and Chamberlain, and Livingston studies).

These ideas for study are given as examples only. Additional areas--varying in the importance of their effects--could be generated also.

At least two lines of research can be identified. One line should help establish the size and direction of effects on group test performance (or on other indicators of impact) resulting from systematic manipulation of the factors listed above. This line of research might be viewed as useful in establishing the validity of sociolinguistic concepts. Such exploratory studies may not have immediate application, but they should prove useful in establishing whether the observed data behave in a way that is consistent with the theory on which the techniques are based.

Another line of research is directed at more specific determination of the effects of applying sociolinguistic techniques to personnel test situations. These effects are reflected in such test statistics as the distribution of item difficulties and in predictive validity coefficients. This approach is consistent with both the goal of changing the alignment of various population groups and the goal of making this alignment more consistent with subsequent performance. To make test language easy at the expense of testing relevant, but difficult, concepts will not be useful. Therefore, in addition to understanding the effects of sociolinguistic manipulations of tests, investigations must also be useful in choosing techniques that will result in more effective personnel selection procedures.

10.3.2. Scientific approach. Social scientists, particularly psychologists, long ago learned that single-factor experiments can lead to confusing and perhaps contradictory results. They are, therefore, aware of the importance of factorial experiments that simultaneously vary several factors. For example, it seems perfectly reasonable to suppose that the results of changing the motivating effect of directions would not be the same for examinees coming from different backgrounds. Specifically, it is hard to imagine that changes in the directions given in a tool knowledge test could be expected to have the same effect on a person enlisting for a medical job and one seeking training in automobile maintenance. Finding the kind and size of any existing difference requires the simultaneous variation of the group tested and the type of directions given.

One can see from the few examples above that the list of possible factors is too long to include each one in a grand factorial experiment; including only two levels of each of the eight factors listed in the previous section would require 256 experimental groups. Conducting such an experiment would be extremely complex, and certainly far beyond anything that has to date proved manageable in the field of personnel testing. A programmatic series of experiments aimed at the systematic development, testing, and application of sociolinguistically-based hypotheses related to test performance seems much more reasonable. This is simply to suggest, in the tradition of scientific practice, that orderly, sequential development and experimentation steps be implemented.

10.3.3. Implementation. The suggested research approach undoubtedly requires a sustained effort. Because of the extensive administration of the current joint services selection test, Armed Services Vocational Aptitude Battery, at the high school level, it would seem that this population (and its subpopulations) would be suitable for research studies for which contracts based on either solicited or unsolicited proposals might be awarded. Most of the data, however, could come from the testing of incumbent military personnel. These data might be efficiently gathered and analyzed by using appropriate experimental designs overlaid on data collection efforts conducted in connection with other military personnel research. In this manner, data might serve the needs of both sociolinguistic and military personnel researchers.

It is difficult to discuss organizational methods to reach a goal so abstract as that of "identifying and developing sociolinguistic principles for application to test construction." It is, therefore, suggested that perhaps teams of specialists composed of sociolinguistic and measurement experts could be allowed to inspect existing personnel tests, be informed about anticipated development efforts, and be encouraged to propose research projects pertinent to the goal at hand. After recommendations are received from those teams and studies completed by them, the most probable areas of development and the most useful organizational arrangements should become clearer. A reasonable immediate outlook is for the development of item evaluation checklists to assure proper and careful attention to good test construction principles, from both a psychometric and a sociolinguistic point of view.

## References

Anastasi, A. *Differential psychology* (2nd ed.). New York: Macmillan, 1958.

Anastasi, A. Culture-fair testing. In N. E. Gronlund (Ed.), *Readings in measurement and evaluation*. New York: Macmillian, 1968.

Angoff, W. H. (Ed.). *The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Test*. New York: College Entrance Examination Board, 1971.

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, 1973, *10*(2), 95-106.

APA Task Force on Employment of Minority Groups. Job testing and the disadvantaged. *American Psychologist*, 1969, *24*(7), 637-649.

*Bakke vs. Regents of the University of California*. California Superior Court, Yolo City, Document No. 31287, March 7, 1975.

Bever, T. G., & Rosenbaum, P. S. Some lexical structures and their empirical validity. In R. A. Jacobs & P. W. Rosenbaum (Eds.), *Readings in English in transformational grammar*. Walthan, Mass.: Ginn and Company, 1970.

Binnick, R. I. *Will* and *Be going to*. In D. Adams et al. (Eds.), *Papers from the Seventh Regional Meeting, Chicago Linguistic Society*. Chicago: Chicago Linguistic Society, 1971.

Binnick, R. I. *Will* and *Be going to* II. In P. M. Peranteau, J. N. Levi, & G. C. Phares (Eds.), *Papers from the Eighth Regional Meeting, Chicago Linguistic Society*. Chicago: Chicago Linguistic Society, 1972.

Boehm, V. R. Negro-White differences in validity of employment and training selection procedures. Summary of research evidence. *Journal of Applied Psychology*, 1972, *56*(1), 33-39.

Bolinger, D. L. Binomials and pitch accept. *Lingua*, 1962, *11*, 34-44.

Bornstein, H., & Chamberlain, K. An investigation of some of the effects of "verbal load" in achievement tests. *American Educational Research Journal*, 1970, *7*(4), 597-604.

Bray, D. W., & Moses, J. L. Personnel selection. *Annual Review of Psychology*, 1972, *23*, 545-576.

Byham, W. C., & Spitzer, M. E. The law and personnel testing. American Management Association, 1971.

Campbell, J. T., Pike, L. W., & Flaugher, R. L. Prediction of job performauce for Negro and White medical technicians: A regression analysis of potential test bias: Predicting job knowledge scores from an aptitude battery (PR-69-6). Princeton, N.J.: Educational Testing Service, 1969.

Canady, H. G. The problem of equating the environment of Negro-White groups for intelligence testing in comparative studies. In R. C. Wilcox (Ed.), The psychological consequences of being a Black American. New York: Wiley, 1971.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. Educational uses of tests with disadvantaged students. American Psychologist, 1975, 30(1), 15-41.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Cronbach, L. J. Five decades of public controversy over mental testing. American Psychologist, 1975, 30(1), 1-14.

Cronbach, L. J. Equity in selection--Where psychometrics and political philosophy meet. Journal of Educational Measurement, 1976, 13(1), 31-42.

Crystal, D., & Quirk, R. Systems of prosodic and paralinguistic features in English. Atlantic Highlands, N.J.: Humanities Press, 1964.

Darlington, R. B. Another look at cultural fairness. Journal of Educational Measurement, 1971, 8, 71-82.

Dyer, H. S. Does coaching help? College Board Review, 1953, 19, 331-335.

Dyer, H. S. A psychometrician views human ability. Teachers College Record, 1960, 61, 394-403.

Einhorn, H. J., & Bass, A. R. Discrimination in employment testing. Psychological Bulletin, 1971, 75, 261-269.

Equal Employment Opportunity Commission. Guidelines in employment selection procedures. Title 29; Chapter XIV, Part 1607. EEOC, 1 August 1970. (Federal Register Document 70-9962.)

Flaugher, R. L. Bias in testing: A review and discussion. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, 1974.

Flesch, R. How to test readability. New York: Harper and Bros., 1951.

French, J. E., & Dear, R. E. Effect of coaching on an aptitude test. Educational and Psychological Measurement, 1959, 19(3), 319-330.

Ginger, A. R. (Ed.). De Funis vs. Odegaard and the University of Washington. Council on Legal Educational Opportunity. Dobbs-Ferry, N.Y.: Oceanic Publications, 1974.

Gordon, D., & Lakoff, G. Conversational postulates. In D. Adams et al. (Eds.), Papers from the Seventh Regional Meeting, Chicago Linguistic Society. Chicago: Chicago Linguistic Society, 1971.

Green, G. M. How to get people to do things with words: The question of imperatives. In R. W. Shuy (Ed.), Some new directions in linguistics. Washington, D.C.: Georgetown University Press, 1973.

Grice, H. P. Logic and conversation. Lecture notes from William James Lectures. Cambridge, Mass.: Harvard University, 1967.

Griggs vs. Duke Power Company. 401 U.S. 424, 1971.

Guion, R. M. Employment tests and discriminatory hiring. Industrial Relations, 1966, 5, 20-37.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. Testing and fair employment. New York: New York Press, 1968.

Lakoff, R. The pragmatics of modality. In P. M. Peranteau, J. N. Levi, & G. C. Phares (Eds.), Papers from the Eighth Regional Meeting, Chicago Linguistic Society. Chicago: Chicago Linguistic Society, 1972.

Lawler, J. Generic to a fault. In P. M. Peranteau, J. N. Levi, & G. C. Phares (Eds.), Papers from the Eighth Regional Meeting, Chicago Linguistic Society. Chicago: Chicago Linguistic Society, 1972.

Livingston, S. A. Verbal overload and achievement tests: A replication. American Educational Research Journal, 1973, 10(2), 155-162.

Lorge, I. Difference or bias in tests of intelligence. In Proceedings of the 1952 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1953.

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture fair selection. Journal of Educational Measurement, 1976, 13(1), 3-30.

Pike, K. L. The intonation of American English. Ann Arbor: University of Michigan Press, 1945.

Reilly, R. R. A note on minority group test bias studies. Psychological Bulletin, 1973, 80, 130-132.

Ruch, F. L. What impact did the Supreme Court decision in the Duke Power case have on employment procedures? Address delivered to the Edison Electric Institute, Industrial Relations Committee, at the Jack Tarr Hotel, San Francisco, May 26, 1971.

Roberts, S. O., & Oppenheim, D. B. The effect of special instruction upon test performance of high school students in Tennessee (ETC RB 66-36 and College Board RDR 66-7, No. 1). Princeton, N.J.: Educational Testing Service, 1966.

Sadock, J. M. Speech act idioms. In P. M. Peranteau, J. N. Levi, & G. C. Phares (Eds.), Papers from the Eighth Regional Meeting, Chicago Linguistic Society. Chicago: Chicago Linguistic Society, 1972.

Samuda, R. J. Psychological testing of American minorities. New York: Dodd, Mean, & Co., 1975.

Schmidt, F. L., Beiner, J. G., & Hunter, J. E. Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 1973, 58, 5-9.

Tannenbaum, A. J. Culture-fair intelligence test. In O. K. Buros (Ed.), Sixth mental measurements yearbook, tests, and reviews. Highland Park, N.J.: Gryphon Press, 1965.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Tyler, L. E. The psychology of individual differences. New York: Meredith, 1965.

U. S. Department of Labor. Manpower report of the President: A report on manpower requirements, resources, utilization, and training. Washington, D. C.: U. S. Government Printing Office, March 1973.

Washington vs. Davis, 96 S. Ct. 2042, U. S. (1976).

Webster's Third International Dictionary. Springfield, Mass.: G. & C. Merriam & Co., 1971.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1978—771-122/86